

Inference on Difference of Means of two Log-Normal Distributions; A Generalized Approach

K. Abdollahnezhad¹, M. Babanezhad², A. A. Jafari³

¹Department of Statistics, Golestan University, Gorgan, Iran

²Department of Statistics, Golestan University, Gorgan, Iran

³Department of Statistics, Yazd University, Yazd, Iran.

Abstract

Over the past decades, various methods for comparing the means of two log-normal have been proposed. Some of them are differing in terms of how the statistic test adjust to accept or to reject the null hypothesis. In this study, a new method of test for comparing the means of two log-normal populations is given through the generalized measure of evidence to have against the null hypothesis. However calculations of this method are simple, we find analytically that the considered method is doing well through comparing the size and power statistic test. In addition to the simulations, an example with real data is illustrated.

Keywords: Generalized p -value; Generalized test variable; Log-normal distribution; Monte Carlo simulation.

1 Introduction

One often encounters with random variables that are inherently positive in some real life applications such as analyzing biological, medical, and industrial data. In this regards the normal distribution is applied in most of applications. In the family of normal distribution the Log normal distribution has a long term applications. In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Further, a variable might be modeled as log-normal if it can be thought of as the multiplicative product of many independent random variables

each of which is positive. The suitability of the log-normal random variable has been investigated by some researchers (Crow and Shimizu, 1988). There are also some recent articles regarding the statistical inference of parameters of several log-normal distributions. For example, one-sided test have been investigated for two distributions with a large sample under the homogeneity of the mean parameters for m log-normal populations (Zhou et al., 1997; Ahmed et al., 2002). Further, exact confidence interval test for the ratio or difference of the means of two log-normal distributions using the generalized variable and generalized p -values through a modified likelihood ratio has been done (Krishnamoorthy and Mathew, 2003; Gill, 2004; Gupta and Li, 2005). In this paper, we consider random samples from two lognormal populations and our interest is to present a test of difference of the means of these two populations. In Section 2, the theory of generalized p -value is introduced. Section 3 is devoted to an exact one-sided test or two-sided test for two log-normal distributions. We compare the size and power of different proposed methods to test of the means of two log-normal populations in Section 4 through simulation. We examine them by a numerical example with real data set. A brief discussion is given in Section 5.

2 Generalized p -value

The concept of generalized p -value was first introduced by Tsui and Weerahandi (1989) to deal with the statistical testing problem in which nuisance parameters are present and it is difficult or impossible to obtain a nontrivial test with a fixed level of significance. The setup is as follows. Let \mathbf{X} be a random variable having density function $f(\mathbf{x}|\boldsymbol{\zeta})$, where $\boldsymbol{\zeta} = (\theta, \boldsymbol{\eta})$ is a vector of unknown parameters, θ is the parameter of interest, and $\boldsymbol{\eta}$ is a vector of nuisance parameters. Suppose we are interested to test

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0, \quad (1)$$

where θ_0 is a specified value.

Let \mathbf{x} denote the observed value of \mathbf{X} and consider a variable $T(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$, by the name of generalized variable. We assume that $T(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$ satisfies the following conditions:

- (i) For fixed \mathbf{x} , the distribution of $T(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$ is free from the nuisance parameters $\boldsymbol{\eta}$.

(ii) $t_{obs} = T(\mathbf{x}; \mathbf{x}, \boldsymbol{\zeta})$ is free from any unknown parameters.

(iii) For fixed \mathbf{x} and $\boldsymbol{\eta}$, $T(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$ is either stochastically increasing or decreasing in θ for any given t .

Under the above conditions, if $T(\mathbf{X}; \mathbf{x}, \boldsymbol{\zeta})$ is stochastically increasing in θ , then the generalized p -value for testing the hypothesis in (1) can be defined as

$$p = \sup_{\theta \leq \theta_o} P(T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\eta}) \geq t^*) = P(T(\mathbf{X}; \mathbf{x}, \theta_o, \boldsymbol{\eta}) \geq t^*), \quad (2)$$

where $t^* = T(\mathbf{X}; \mathbf{x}, \theta_o, \boldsymbol{\eta})$.

For further details and for several applications based on the generalized p -value, we refer to the book by Weerahandi (1995).

3 A generalized test variable

Let $Y_{ij} = \ln(X_{ij}) \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$, $j = 1, 2, \dots, n_i$ be independent random samples from two log-normal populations. We know that $M_i = E(X_{ij}) = \exp(\mu_i + 0.5\sigma_i^2)$. The problem of our interest is one sided and two sided test hypothesis about $\eta = M_1 - M_2$.

In this section, using the concept of generalized p -value, we test

$$H_o : M_1 \leq M_2 \quad vs \quad H_1 : M_1 > M_2, \quad (3)$$

which is equivalent to

$$H_o : \theta \leq 0 \quad vs \quad H_1 : \theta > 0, \quad (4)$$

where $\theta = \ln M_1 - \ln M_2$.

The MLE's for μ_i and σ_i^2 ($i = 1, 2$) are \bar{Y}_i and S_i^2 , respectively, where

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad , \quad S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

Now, consider

$$\begin{aligned} T &= \bar{y}_{1.} - \bar{y}_{2.} + \frac{\bar{Y}_{2.} - \bar{Y}_{1.} - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sqrt{\frac{\sigma_1^2 s_1^2}{n_1 S_1^2} + \frac{\sigma_2^2 s_2^2}{n_2 S_2^2} + \frac{\sigma_1^2 s_1^2}{2S_1^2} - \frac{\sigma_2^2 s_2^2}{2S_2^2}} - \theta \\ &= \bar{y}_{1.} - \bar{y}_{2.} + Z \sqrt{\frac{s_1^2}{U_1} + \frac{s_2^2}{U_2}} + \frac{n_1 s_1^2}{2U_1} - \frac{n_2 s_2^2}{2U_2} - \theta, \end{aligned}$$

where

$$Z = \frac{\bar{Y}_2 - \bar{Y}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1),$$

and

$$U_i = \frac{n_i S_i^2}{\sigma_i^2} \sim \chi_{(n_i-1)}^2, \quad i = 1, 2,$$

are three independent random variables, and \bar{y}_i and s_i^2 are observed values of \bar{Y}_i and S_i^2 , respectively. Then, T is a generalized variable for θ because

- i) $t_{obs} = 0$
- ii) distribution of T is free from the nuisance parameters μ_i and σ_i^2 .
- iii) the distribution of T is an increasing function with respect to θ .

Thus the generalized p -value for the null hypothesis (3) is given by

$$p = P(T \leq t_{obs} | \theta = 0) = E\left(\Phi\left(\frac{\bar{y}_2 - \bar{y}_1 + \frac{n_2 s_2^2}{2U_2} - \frac{n_1 s_1^2}{2U_1}}{\sqrt{\frac{s_1^2}{U_1} + \frac{s_2^2}{U_2}}}\right)\right), \quad (5)$$

where $\Phi(\cdot)$ is the standard normal distribution function and the expectation is taken with respect to independent chi-square random variables, U_1 and U_2 .

This generalized p -value can be well approximated by a Monte Carlo simulation using the following algorithm:

Algorithm 1. For a given data set x_{i1}, \dots, x_{in_i} , set $y_{ij} = \ln(x_{ij})$, $i = 1, \dots, k$, $j = 1, 2$.

Compute $\bar{y}_1, \bar{y}_2, s_1^2, s_2^2$

For $l = 1$ to m

Generate $U_1 \sim \chi_{(n_1-1)}^2$, $U_2 \sim \chi_{(n_2-1)}^2$.

Calculate $T_l = \Phi\left(\frac{\bar{y}_2 - \bar{y}_1 + \frac{n_2 s_2^2}{2U_2} - \frac{n_1 s_1^2}{2U_1}}{\sqrt{\frac{s_1^2}{U_1} + \frac{s_2^2}{U_2}}}\right)$.

$\frac{1}{m} \sum_{l=1}^m T_l$ is a Monte Carlo estimate of generalized p -value for the null hypothesis (3).

The generalized p -value in (5) is used for one sided test hypothesis but we can use this generalized p -value for two sided test hypothesis by

$$p = 2 \min\{p, 1 - p\},$$

where p is the generalized p -value in (5).

4 Simulation Study

To investigate the power of the considered test statistics in finite samples, we conducted a simulation experiment. To do so, several data set from two log- normal distributions with $\mu_2 = 0$ were generated. For each scenarios 10000 sample size are performed. The size and the power of the considered test statistics are summarized in tables in table 2 and 3. These tests are (a) generalized p -value in (5) (b) generalized p -value by Krishnamoorthy and Mathew (2003) (c) Z-score test by Zhou et al. (1997). The simulation study indicates that (i) The size for (a) and (b) are close to 0.05 and the powers are close to each other. (ii) The size of (c) is very larger than nominal level, 0.05.

5 Numerical examples

The data show the amount of rainfall (in acre-feet) from 52 clouds; 26 clouds were chosen at random and seeded with silver nitrate. We can show that log-normal model fits the data. The summary statistics for the log-transformed data are given in Table 1.

Table 1: The summary statistics for the log-transformed data of rainfall

Clouds	n_i	\bar{y}_i	s_i^2
seeded clouds	26	5.134	2.46
unseeded clouds	26	3.990	2.60

In order to understand the effect of silver nitrate seeding, we like to test

$$H_o : M_1 = M_2 \text{ vs } H_1 : M_1 > M_2, \quad (6)$$

where $M_i = \exp(\mu_i + 0.5\sigma_i^2)$, $i = 1, 2$.

The p -values for our generalized approach, Krishnamoorthy and Mathew approach and Z-score test are 0.0779, 0.0747 and 0.0599 respectively. Therefore, we cannot reject H_o at the level of 0.05, using all 3 methods.

References

- [1] Ahmed, S. E., Tomkins, R. J. and Volodin, A.I. (2001). Test of homogeneity of parallel samples from lognormal populations with unequal variances, *Journal of Statistical Research*, 35, no 2, 25-33.
- [2] Crow, E. L. and Shimizu, K. (1988). *Lognormal distribution*, Marcel Dekker: New York.
- [3] Gill, P. S. (2004). Small sample inference for the comparison of means of lognormal distribution, *Biometrics*, 60, 525-527.
- [4] Gupta, R. C. and Li, X. (2005). Statistical inferences on the common mean of two lognormal distributions and some applications in reliability, appeared in *Computational Statistics and Data Analysis*.
- [5] Krishnamoorthy, K. and Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence interval, *Journal of Statistical Planning and Inference*, 115, 103-121.
- [6] Krishnamoorthy, K. and Yong Lu. (2003). Inferences on the common mean of several normal populations based on the generalized variable method, *Biometrics*, 59, 237-247.
- [7] Tsui, K. W. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypothesis in the presence of nuisance parameters, *J. Am. Statist. Assoc.*, 84, 602-607.
- [8] Weerahandi, S. (1993). Generalized confidence intervals, *J. Am. Statist. Assoc.*, 88, 899-905.
- [9] Weerahandi, S. (1995a). *Exact statistical methods for data analysis*, Springer, New York.
- [10] Weerahandi, S. and Berger, V. W. (1999). Exact inference for growth curves with interclass correlation structure, *Biometrics*, 55, 921-924.
- [11] Zhou, X. H., Gao, S. and Hui, S.L. (1997). Methods for comparing the means of two independent lognormal samples, *Biometrics*, 53, 1129-1135.

- [12] Zhou, X. H. and Tu. W. (1999). Comparison of several independent population means when their samples contain lognormal and possibly zero observations, *Biometrics*, 55, 645-651.

Table 2: Simulated sizes of the tests at 5% significance level when $\mu_2 = 0$.

n_1	n_2	μ_1	σ_1^2	σ_2^2	(a)	(b)	(c)
4	4	1	2	4	421	436	1091
		0	3	3	344	405	367
		5	2	12	464	510	2168
10	10	0	12	12	392	391	112
		1	2	4	612	603	895
		0	3	3	546	581	432
		5	2	12	515	552	1433
		0	12	12	538	538	386
		0	1	1	512	524	614
25	25	0	5	5	521	522	516
		0	10	10	486	531	446
		0	100	100	521	531	396
		2	4	8	538	520	828
		4	8	16	492	493	851
		0	1	1	391	467	506
40	25	0	5	5	412	425	491
		0	10	10	459	416	512
		0	1	1	382	376	364
25	40	0	5	5	394	373	244
		0	10	10	435	412	199
		5	2	12	521	510	1061
25	40	5	2	12	312	341	586
40	40	8	4	20	536	492	932
		14	4	32	513	546	922
100	25	0	1	1	451	473	664
		0	5	5	374	379	714
		0	10	10	396	388	720
25	100	0	1	1	482	464	295

Table 3: Simulated powers of the tests at 5% significance level when $\mu_2 = 0$.

n_1	n_2	μ_1	σ_1^2	σ_2^2	(a)	(b)	(c)
4	4	0	12	4	1523	1496	364
		3	2	4	1261	1204	3832
		0	20	4	2610	2601	334
		4	1	1	5753	5772	9621
10	10	0	12	4	4136	4089	2334
		0	20	4	6941	6903	4562
		3	2	4	2961	3114	5173
		4	1	1	9931	9942	9990
25	25	1	1	1	8370	8345	8917
		1	5	5	1916	1843	2081
		0	4	2	3564	3521	3157
		1	10	10	1126	1123	1250
		0	9	7	1314	1360	1225
		0	4	1	7411	7390	6854
40	25	1	1	1	8392	8324	9036
		1	5	5	2023	2025	2736
		1	10	10	1173	1123	1492
25	40	1	1	1	9194	9135	9401
		1	5	5	2243	2307	2159
		1	10	10	1120	1145	784
40	25	1	5	4	8836	8814	4649
		1	10	9	1831	1734	2263
25	40	1	5	4	4464	4482	4955
		1	10	9	1956	1893	1493
100	25	1	1	1	8834	8762	9512
		1	5	5	1856	1932	3364
		1	10	10	942	913	1825
25	100	1	1	1	9984	9913	9893